Lattice Modeling: Accuracy of Energy Calculations

BORIS A. REVA,*,† MICHEL F. SANNER,* and ARTHUR J. OLSON*,‡

**Department of Molecular Biology, The Scripps Research Institute, 10666 North Torrey Pines Road, California 92037*

†Institute of Mathematical Problems of Biology, Russian Academy of Sciences, 142292, Pushchino, Moscow Region, Russian Federation

ALEXEI V. FINKELSTEIN

Institute of Protein Research, Russian Academy of Sciences, 142292, Pushchino, Moscow Region, Russian Federation

Received 16 May 1995; accepted 25 August 1995

ABSTRACT

Energy calculations based on lattice models of protein chains are always approximate, because any such a model distorts distances between chain links and, consequently, the energies of interaction between them. The energetic errors of lattice models are examined here for 15 proteins of different sizes and types of secondary structure, for lattice spacings ranging from 0.25 to 2.5 Å. The lattice models are derived using previously described algorithms which insure a minimal root mean square (rms) deviation from the off-lattice structure for any given lattice–protein orientation. For each protein structure we computed a set of different lattice models with virtually equal rms deviations, and then compared their energies. Energy calculations were based on the pairwise potentials. We found that the energies of lattice models follows a normal distribution with a nonnegligible dispersion, even at a fine lattice spacing of 0.25 Å. For any lattice model of a protein, the lattice spacing must be 1.0 Å or less in order to be able to distinguish energetically between the folded and extended states. However, when an ensemble of lattice models is considered, this distinction can be made for lattice spacing up to 2.0 Å. We conclude that to attain a better approximation of the protein lattice model energies, one must adjust potentials to the lattice spacing.

‡Author to whom all correspondence should be addressed.

Journal of Computational Chemistry, Vol. 17, No. 8, 1025–1032 (1996)

CCC 0192-8651 / 96 / 081025-08

Introduction

The main advantage of lattice models of polymers (Fig. 1) (where each chain link is embedded in the points of a 3-dimensional grid) is that these models have a finite number of conformations, which enables an extensive exploration of their conformational space. Lattice modeling is used today in virtually all theoretical studies of protein folding, kinetics, and thermodynamics.[4-7]

Models based on fine lattices can reproduce the energies of off-lattice structures to any given accuracy. However, these fine lattice models negate the computational advantages of lattice modeling since the number of conformations to be explored approaches the continuous model. The question arises, what is the coarsest lattice which still gives a reasonable approximation of the energy of the actual, off-lattice structure.

In this article we concentrate only on the errors in nonbonded energy estimation which arise from lattice approximations, not on those coming from the errors and uncertainties in the off-lattice energetic parameters; these other errors have been discussed elsewhere.[8-10]

The energies of protein structures and of their lattice models are different because the models distort the positions of the structural links. The distances between them become slightly smaller or larger as compared with those in the actual off-lattice structure.

Thus, all energy calculations based on lattice models are inevitably approximate, as are the corresponding energy landscapes (see Fig. 2).

In particular this means that the minimal energy structure can be approximated, with a small root mean square (rms) deviation, by a lattice model of high energy; and some other structure can be approximated by a lattice model of low energy.

To apply lattice models for energy computations (in particular for searching for the stable state

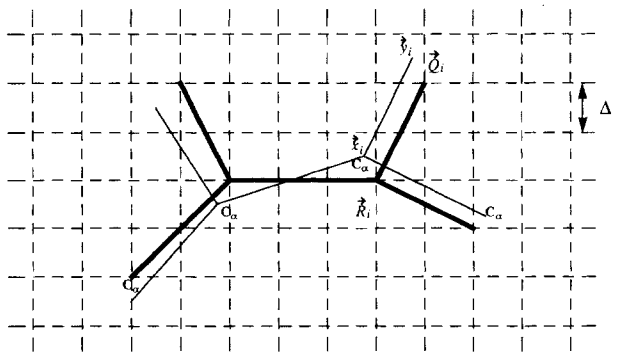

FIGURE 1. Lattice modeling of protein molecules with side chains. A two-dimensional case is shown for simplicity. A fragment of a protein structure is shown by the thin line, its lattice model by the solid line, and the lattice itself is shown by dashed lines. The main chain is represented by the positions of C_α atoms; the side chains are shown by segments connecting C_α atoms and the centers of their corresponding residues (for glycine the C_α-atom position is also the center of the side chain); $\mathbf{x}_i$ is the actual off-lattice position of the C_α atom of residue i and $\mathbf{R}_i$ is its position on the lattice; $\mathbf{y}_i$, $\mathbf{Q}_i$ are actual and lattice coordinates of the center of the side chain of residue i. The algorithm for building a lattice model chooses that combination of lattice points which minimizes the deviation $\Sigma(\mathbf{x}_i - \mathbf{r}_i)^2 + \Sigma(\mathbf{y}_i - \mathbf{Q}_i)^2$ under the conditions $||\mathbf{R}_i - \mathbf{R}_{i+1}| - |\mathbf{x}_i - \mathbf{x}_{i+1}|| \leq \gamma$, which maintains the chain connectivity, and $||\mathbf{Q}_i - \mathbf{R}_i| - |\mathbf{y}_i - \mathbf{x}_i|| \leq \gamma$, which binds the side chain to its C_α atom. In this study, $\gamma = \Delta/2$ where Δ is the lattice spacing.

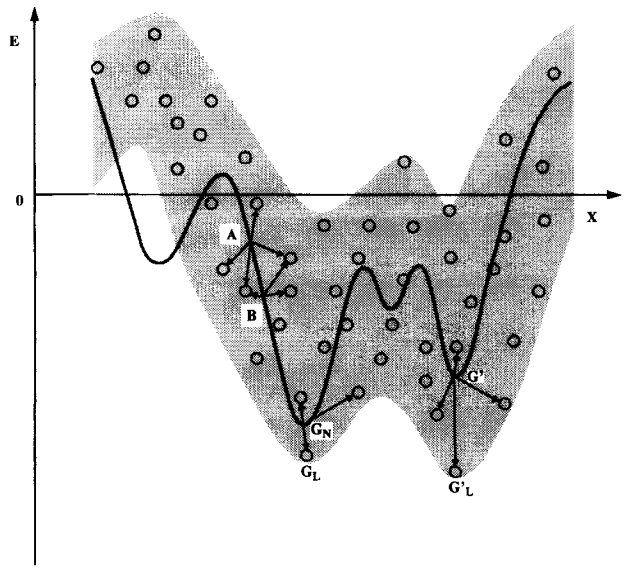

FIGURE 2. The energy landscape of a protein molecule (one-dimensional cross section is shown schematically by a solid line) is reproduced approximately by a variety of lattice models (circles in shadow area). Each point on the solid curve corresponds to a particular 3-dimensional conformation of the protein and its particular energy. Each of these off-lattice conformations (e.g., points A, B) can be approximated, as is shown by arrows, by a number of lattice models with roughly equal rms deviations but with different energies. Thus, the energy of the lattice model G_L (approximating the native fold G_N actually having the minimal energy) can be higher than the energy of the lattice model G'_L approximating some other fold G'.

of a chain) it is necessary to know: 1) What are the reliable differences in the lattice model energies? And 2) how do they depend upon the spacing, the length of the protein molecule, and the energetic parameters?

A direct way to answer these questions is to consider a number of protein conformations, construct their corresponding lattice models, and then compare their energies. We compare energies of native protein folds with energies of the same chains in extended conformations.

Methods

The native folds of 15 protein chains (Table I) were taken from the Protein Data Bank.[11]

The extended conformations of the protein chains were obtained by assigning β-strand parameters to all ϕ, ψ angles of the main chain ($-120°$, $135°$, respectively), and $180°$ for χ angles of the side chains.

To obtain the lattice models (Fig. 1) we used our recently developed algorithms[1,2] which guarantees, for any given chain–lattice orientation, the lattice model with the minimal rms deviation.

For both off-lattice conformations of the protein chain (i.e., the native fold and extended conformation), we built a lattice model corresponding to each of 100 randomly chosen lattice–chain orientations.

Energy calculations were based on the force-field parameters of pairwise side chain–side chain interactions.[3]

According to this force field, the energy of long-range interactions E_{L-R} has the following form:

$$E_{L-R} = \frac{1}{2}\sum_{i=1}^{N-1}\sum_{\substack{j=1\\ j\neq i}}^{N} \nu_{i,j}(\alpha_i, \mathbf{r}_i, \alpha_j, \mathbf{r}_j), \qquad (1)$$

where α_i, α_j are types, and $\mathbf{r}_i$, $\mathbf{r}_j$ are position of the centers of the side chains of residues i and j, respectively. The pairwise potential is defined as:

$$\nu_{i,j}(\alpha_i, \mathbf{r}_i, \alpha_j, \mathbf{r}_j) = \begin{cases} 0, & \text{if } |\mathbf{r}_i - \mathbf{r}_j| > P_1(\alpha_i, \alpha_j) \\ V(\alpha_i, \alpha_j), & \text{if } P_1(\alpha_i, \alpha_j) \le |\mathbf{r}_i - \mathbf{r}_j| \le P_2(\alpha_i, \alpha_j) \\ +4RT, & \text{if } |\mathbf{r}_i - \mathbf{r}_j| < P_2(\alpha_i, \alpha_j), \end{cases} \qquad (2)$$

where energy V and distances P_1, P_2 are obtained from protein statistics, and the value $+4RT$ is used for residues in sterically unfavorable positions.[3]

TABLE I.
List of Proteins.

Protein Name	PDB Code	No. Residues	Structural Class
Epidermal growth factor	1epg	53	Small proteins
Ovomucoid III domain from silver pheasant	2ovo	56	Small proteins
Eglin C (α-chymotrypsin inhibitor)	1acb	63	$\alpha + \beta$
α-Cobratoxin (cobra)	2ctx	71	Small proteins
Parvalbumin	1cdp	79	α
Subtilisin inhibitor	3sic	107	$\alpha + \beta$
FK-506 binding protein	1fkf	107	$\alpha + \beta$
Thioredoxin (*Escherichia coli*)	2trx	108	α / β
Hemerythrin	1hmd	113	α
Snake phospholipase A2	1ppa	121	α
Ribonuclease A from bovine	1rat	124	$\alpha + \beta$
Azurin	2aza	129	β
Fatty acid binding protein	1ifb	131	β
Flavodoxin (*Chondrus crispus*)	2fcr	173	α / β
Immunoglobulin (H chain, Fab D1.3)	1fdl	218	β

TABLE II.
Energy Characteristics of Different Lattice Approximations

		Lattice Spacing 2.5 Å						Lattice Spacing 1.0 Å						Lattice Spacing 0.25 Å					
Protein	ΔE	⟨rms⟩	$\langle \Delta E \rangle_L$	SD	ΔE_{min}	C_{fold}	C_{ext}	⟨rms⟩	$\langle \Delta E \rangle_L$	SD	ΔE_{min}	C_{fold}	C_{ext}	⟨rms⟩	$\langle \Delta E \rangle_L$	SD	ΔE_{min}	C_{fold}	C_{ext}
1epg	−63.2	1.38	−39.4	20.2	−60.9	−0.02	−0.07	0.52	−54.8	15.6	−75.9	0.05	−0.14	0.13	−66.5	5.18	−72.3	−0.02	−0.04
2ovo	−76.9	1.38	−3.34	20.2	−28.5	−0.01	0.11	0.53	−56.8	13.2	−76.2	0.00	0.11	0.13	−75.5	3.81	−81.8	−0.11	−0.12
1acb	−64.9	1.37	−10.8	19.8	−37.3	0.13	−0.09	0.53	−57.3	10.3	−67.1	−0.03	0.01	0.13	−62.6	4.65	−67.3	−0.13	0.08
2ctx	−81.2	1.39	−23.7	26.2	−49.7	−0.03	−0.04	0.52	−60.5	20.2	−96.5	0.16	0.04	0.13	−81.9	8.42	−99.7	−0.03	−0.07
1cdp	−61.8	1.37	30.7	31.1	−12.7	0.12	0.08	0.53	−57.3	10.3	−67.1	−0.03	0.01	0.13	−59.5	7.39	−64.8	0.17	0.22
3sic	−75.4	1.39	30.1	29.0	4.20	0.19	0.17	0.52	−47.9	17.1	−64.2	0.13	−0.05	0.13	−72.5	6.92	−83.0	−0.15	0.04
1fkf	−96.1	1.38	19.9	27.9	3.70	0.12	0.14	0.53	−76.6	14.6	−94.3	−0.06	0.06	0.13	−94.7	6.37	−101.	0.05	0.14
2trx	−75.0	1.37	25.3	25.0	7.90	0.13	0.06	0.52	−61.8	15.8	−83.3	−0.01	−0.09	0.13	−79.3	7.23	−87.0	0.17	−0.05
1hmd	−69.1	1.38	10.0	26.3	−18.7	0.09	−0.13	0.53	−58.2	15.5	−75.1	0.14	−0.04	0.13	−67.4	6.80	−78.8	0.23	0.03
1ppa	−85.3	1.38	28.0	28.2	0.90	0.10	0.07	0.53	−46.4	23.9	−78.9	0.02	−0.01	0.13	−92.2	8.55	−98.2	0.09	0.04
1rat	−91.0	1.39	50.7	30.5	26.5	0.05	0.11	0.53	−53.2	18.8	−94.1	0.23	−0.01	0.13	−86.2	8.41	−97.2	−0.11	−0.10
2aza	−93.3	1.38	46.4	28.2	22.4	0.12	−0.03	0.53	−46.7	21.5	−72.1	0.19	−0.05	0.13	−86.6	8.71	−93.9	0.08	−0.13
1ifb	−83.4	1.37	22.9	28.4	2.90	0.18	−0.07	0.53	−72.2	16.4	−89.7	0.13	0.07	0.13	−89.1	8.56	−105.	0.06	−0.20
2fcr	−103.	1.38	42.2	31.6	−1.20	0.00	−0.11	0.53	−92.1	19.4	−117.	−0.02	0.04	0.13	−99.8	10.0	−119.	0.10	−0.04
2fdl	−85.2	1.40	104.	36.3	70.1	0.08	0.07	0.53	−44.7	20.8	−70.5	0.06	−0.04	0.13	−77.1	11.4	−93.1	−0.06	0.02

ΔE is the energy of the folded structure relative to the extended one; $\langle \Delta E \rangle_L$ is the mean energy of the folded lattice models relative to the mean energy of the extended ones; ⟨rms⟩ is the mean rms deviation of lattice models from the native coordinates; SD is a standard deviation of the relative energy from its mean value $\langle \Delta E \rangle_L$; ΔE_{min} is the difference between the energies of the lowest energy models of the folded and extended states; C_{fold} and C_{ext} are the correlation between energies and rms deviations for lattice models of the folded and extended conformations, respectively.

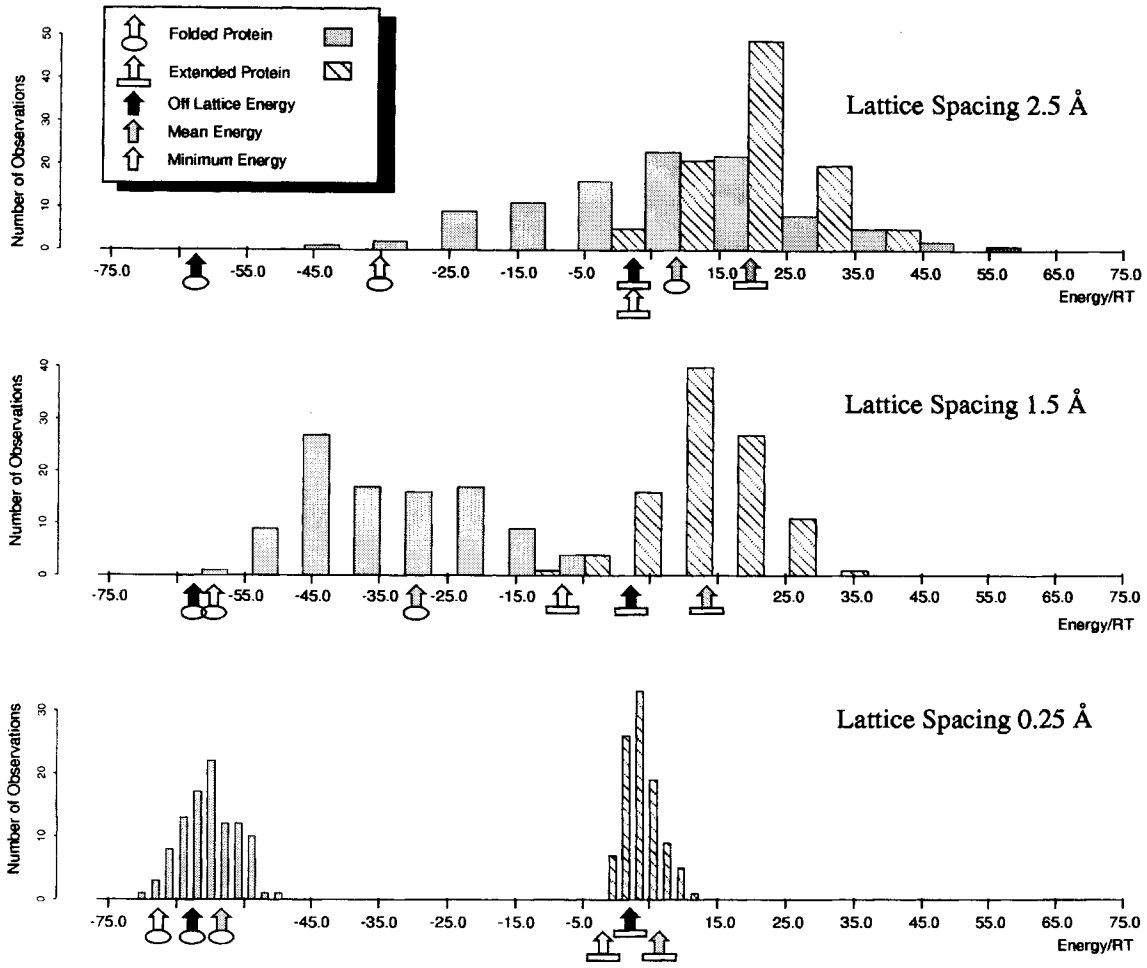

FIGURE 3. Typical histograms of the energy distribution for lattice models of folded and extended conformations of α-chymotrypsin inhibitor (1acb). Each histogram is built for 100 models corresponding to different, randomly chosen chain–lattice orientations. The off-lattice, mean lattice, and minimum lattice energies are indicated with arrows for both folded and extended conformations. Grey bars show the energy distribution of the 100 lattice models of the folded structure; striped bars show the same for the extended structure.

Having these potentials, we computed the off-lattice energies of all 15 protein chains in folded and extended conformations, and found their differences $\Delta E = E_{fold} - E_{ext}$. Then we computed the energies of all lattice models approximating the native and the extended chain conformations, found (averaging over 100 lattice models) the mean difference $\langle \Delta E \rangle_L = \langle E_{fold} \rangle_L - \langle E_{ext} \rangle_L$ in energies of the models, and the standard deviation (SD) of this energy differences (since we had 100 models of native folds and 100 models of extended chains, the number of the differences is 10,000 for each protein). We also calculated $\Delta E_{min} = \min\{E_{fold}\}_L - \min\{E_{ext}\}_L$, the difference between the lowest energy lattice model approximating the native fold, and the lowest energy model approximating the extended state of the chain.

Results and Discussion

The results of computations are given in Table II and in Figures 3–5. The data show that lattice models introduce a considerable error in protein molecule energy calculations, even for very fine lattice spacings (less than 1 Å).

Energies of the lattice models follow a normal distribution (see Fig. 3 for example). The devia-

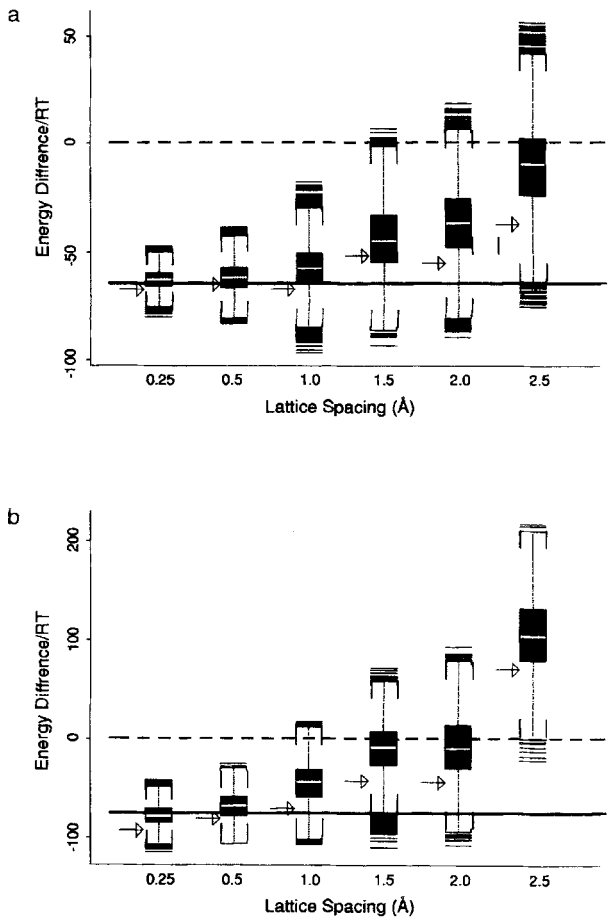

FIGURE 4. (A) Distribution of differences between the energies of 100 lattice models of native fold of α-chymotrypsin inhibitor (1acb), and of 100 lattice models of its extended chain (each spectrum contains 100 * 100 = 10,000 lines) for six different lattice spacings. (Assuming a Gaussian distribution, 99.3% of the data falls inside the brackets.) The short horizontal white lines in the interior of the boxes corresponds to the $\langle \Delta E \rangle_L$ value; the arrows correspond to ΔE_{min} values. The solid horizontal line indicates the actual energy difference between the off-lattice native and extended conformations; the dotted line corresponds to zero energy difference. (B) The same as in (A) for the immunoglobulin (1fdl).

tions of the mean values from the corresponding off-lattice energies and the dispersions depend principally upon the lattice spacing and not significantly upon the size of the molecule or its secondary structure type (Table II).

Table II shows also that there is no correlation between rms deviations and energies of lattice models approximating protein molecule structures at any given lattice spacing (columns C_{fold} and C_{ext}). This demonstrates that at any given spacing the goodness of the model's geometric fit does not imply a good energy approximation.

The test we used for estimation of the errors of lattice models (i.e., a comparison of energies between models of folded and unfolded states of a protein) gives us "an upper limit estimate" for lattice spacing to be used for protein folding simulations. This criterion was applied for three different cases: we compared energy differences for randomly chosen lattice models, for mean energies of models, and for minimal energy models.

Beyond a lattice spacing of 1.0 Å, the distribution of errors is so broad that the energy difference between randomly chosen models of folded and extended conformations says nothing about the actual difference between their energies. One can see this also from the values of correlations between off-lattice and lattice energy differences with the corresponding lattice spacings given in parentheses: −0.311 (2.5 Å), −0.134 (2.0 Å), 0.153 (1.5 Å), 0.361 (1.0 Å), 0.714 (0.5 Å), 0.764 (0.25 Å).

However, the differences between average energies and between the minimal energies of native and extended lattice models are more reliable. These differences are capable of discriminating folded from extended structures up to a lattice spacing of 1.5 and 2.0 Å (Figs. 4, 5).

Protein folding predictions depend upon the accuracy of estimations of energy differences between different states of the protein molecule. Our results [correlations −0.529 (2.5 Å), 0.369 (2.0 Å), 0.314 (1.5 Å), 0.718 (1.0 Å), 0.849 (0.5 Å), 0.917 (0.25 Å)] show that the reliability of differences in minimal energies (see also the Fig. 4) favors the correspondence between the low-energy folds achieved on fine lattices to the off-lattice one.

However, no algorithm will be able to choose between two folds with an energy difference less than the value of the lattice error. The values of the SD reported here (SD columns of Table II) gives an estimate of the differences in energies which can be resolved using difference lattice spacing.

Conclusions

For most of the proteins studied, a comparison of lattice models which approximate the geometry of the two competing conformations (folded and extended) with equal accuracy, gives no informa-

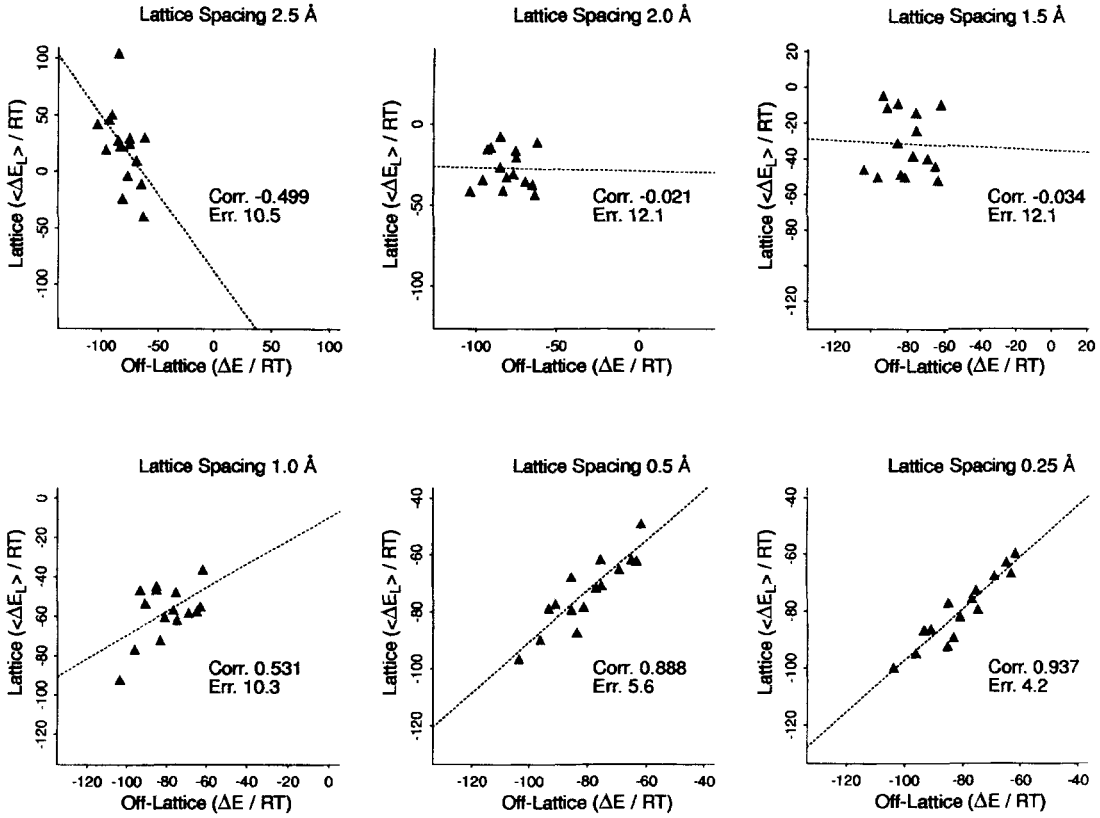

FIGURE 5. Correlation between ΔE, the energy difference between folded and extended conformations, and $\langle \Delta E \rangle_L$, the mean energy difference for their lattice models. The averaging is done over 100 lattice models of the native and 100 models of the extended conformations. The triangles correspond to the proteins studied (Table I). For each of the lattice spacings the correlation coefficient (corr) and the least square root deviation (err) are given.

tion on the actual energy of these conformations for a spacing greater than 1.0 Å; the use of ensemble information from the lattice models (either mean energies or minimal energies) can increase the reliability of energy differences in coarser lattices (up to 2.0 Å).

These observations suggest that a more extensive sampling of lattice models around any potential energy minimum should be used in establishing a global minimum.

It seems that the main reason for the inaccuracy of energy calculations done with lattice models is that the potentials of link–link interactions are too rigid, especially at small distances; and small change of link–link distances can greatly change the energy of interactions of the links. This suggests that one can adjust potentials to the spacing of a given lattice, as previously done for lattice approximations of covalent bonding.[12] The results of such adjustment of potentials will be discussed in a subsequent article.

Acknowledgments

This work was supported by NIH Grant PO1GM38794 (to A.J.O.). The authors are grateful to J. Skolnick and A. Kolinski for providing parameters and valuable discussions. A.V.F. acknowledges the financial support on the Russian Foundation for Basic Research (Grant 93-04-6636). This is publication 9216-MB of The Scripps Research Institute.

References

1. B. A. Reva, D. S. Rykunov, A. J. Olson, and A. V. Finkelstein, *J. Comp. Biol.*, **2**, 527 (1995).
2. D. S. Rykunov, B. A. Reva, and A. V. Finkelstein, *Proteins*, **22**, 100 (1995).
3. A. Kolinski, A. Godzik, and J. Skolnick, *J. Chem. Phys.*, **98**, 1 (1993).

4. D. Covell and R. Jernigan, *Biochemistry*, **29**, 3287 (1990).
5. A. Kolinski and J. Skolnick, *Proteins*, **18**, 338 (1994).
6. A. Kolinski and J. Skolnick, *Proteins*, **18**, 353 (1994).
7. D. Hind and M. Levitt, *J. Mol. Biol.*, **243**, 668 (1994).
8. J. D. Bryngelson, *J. Chem. Phys.*, **100**, 6038 (1994).
9. A. V. Finkelstein, A. M. Gutin, and A. Ya. Badretdinov, *Proteins*, **23**, 151 (1995).
10. A. V. Finkelstein, A. M. Gutin, and A. Ya. Badretdinov, *Subcellular Biochem.*, **34**, 1 (1994).
11. F. C. Bernstein, T. F. Koetzle, E. F. JrMeyer, M. D. Brice, O. Kennard, T. Shimanouchi, and M. Tasumi, *J. Mol. Biol.*, **112**, 535 (1977).
12. A. V. Finkelstein and B. A. Reva, *Protein Eng.* **5**, 617 (1992).